Weekly Report

April 28, 2019

1 Work

1. 本周在进行unpair setting下的图片增强，目前还在测试思路的可能性。
2. 完成了VAST 2019 审稿
3. 完成了图布局论文其他算法的性能测试
4. 工作时长：工作日每天9个小时，周末共10个小时，共55个小时。

1.1 工作进度

Table 1: 工作进度

项目	进度	截止时间
DRGraph	完成了其他算法的性能测试	
unpair 低光照图片增强	目前初步的实验效果不佳	
NIPS	Adversarial Attack	2019.5.23

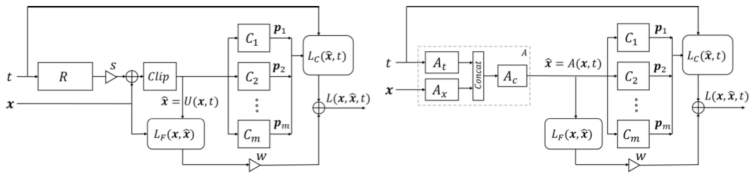
2 Paper Reading

2.1 UPSET and ANGRI : Breaking High Performance Image Classifiers

文章提出两种方案生成对抗案例，1）UPSET为每一个类生成一个对应的对抗扰动，2）ANGRI为每一张图片生成单独的对抗扰动

2.2 Houdini: Fooling Deep Structured Prediction Models

大部分对抗攻击的目标是预测的类别和真实类别不同，这有时候比较难以优化，Houdini 转而采用预测出来各个类别值的大小来控制loss，从而我们可以用于不同任务的攻击。



(a) Training scheme for UPSET(U). (b) Training scheme for ANGRI(A).

Figure 1: #1

$$\bar{\ell}_H(\theta, x, y) = \mathbb{P}_{\gamma \sim \mathcal{N}(0,1)}\Big[g_\theta(x, y) - g_\theta(x, \hat{y}) < \gamma\Big] \cdot \ell(\hat{y}, y)$$

Figure 2: #2

2.3 Art of singular vectors and universal adversarial perturbations

加上对抗扰动的影响可以被梯度所量化，只要使得扰动和梯度之间的关系可以最大化就可以保证最小扰动造成最大的影响。

Let us denote the outputs of the i-th hidden layer of the network by $f_i(x)$. Then for a small vector ε we have

$$f_i(x + \varepsilon) - f_i(x) \approx J_i(x)\varepsilon,$$

where

$$J_i(x) = \frac{\partial f_i}{\partial x}\bigg|_x,$$

is the Jacobian matrix of f_i. Thus, for any q-norm

$$\|f_i(x + \varepsilon) - f_i(x)\|_q \approx \|J_i(x)\varepsilon\|_q, \tag{4}$$

We can conclude that for perturbations which are small in magnitude in order to sufficiently perturb the output of a hidden layer, it is sufficient to maximize right-hand side of the eq. (4). It seems reasonable to suggest that while propagating further in the network it will dramatically change the predicted label of x.

Thus to construct an adversarial perturbation for an individual image x we need to solve

$$\|J_i(x)\varepsilon\|_q \to \max, \quad \|\varepsilon\|_p = L, \tag{5}$$

Figure 3: #3

2.4 Boosting Adversarial Attacks with Momentum

Momentum iterative fast gradient sign method 作为对iterative fast gradient sign method的扩展，它使用了动量项来保证梯度的准确性和稳定性，从而达到更好的攻击效果。